Séminaire de ***M***éthodologie ***S***tatistique

Mercredi 30 novembre 2016

9h30-12h30, Insee - Malakoff 1 - salle Malinvaud

Big Data et Statistiques Publiques : questions de méthodes

La prolifération exceptionnelle de données, souvent désignées sous le terme de « Big Data », est parfois considérée comme une opportunité pour améliorer et enrichir la production d'information. En parallèle, les technologies permettant de traiter des données très volumineuses ou de format complexe se sont développées à un rythme rapide sur la période récente et fournissent de nouveaux outils aux statisticiens. L'utilisation de ces nouvelles données représente cependant un enjeu technique et statistique dont le praticien doit avoir une bonne compréhension pour faire des choix méthodologiques raisonnés.

Ce séminaire sera l'occasion de présenter un aperçu des expérimentations menées sur ces thèmes au sein du département des méthodes statistiques. Après une introduction générale qui présentera le contexte, notamment européen, de l'exploitation de nouvelles sources pour la statistique publique, les exposés aborderont des bilans pratiques de ces expérimentations. Tout d'abord, un premier exposé présentera les solutions techniques et logicielles qui peuvent s'avérer nécessaires pour traiter de très grands volumes de données. Un deuxième exposé proposera une introduction aux méthodes d'apprentissage automatique (« machine learning »), au travers en particulier d'une exploitation de données de téléphonie mobile. Enfin, un dernier exposé fournira une présentation pratique des outils qui peuvent être mobilisés pour traiter des données de formats moins standards (texte ou réseau).

Des Big Data pour la statistique publique ? Pourquoi et comment ?

Pauline Givord – *Division Méthodes appliquées de l'économétrie et de l'évaluation, Insee*

Les outils du Big Data vu par un statisticien

Benjamin Sakarovitch – *Division Méthodes appliquées de l'économétrie et de l'évaluation, Insee*

Les méthodes de machine learning, un investissement utile pour la statistique publique

Stéphanie Combes – *Division Méthodes appliquées de l'économétrie et de l'évaluation, Insee*

Comment exploiter des données aux formats de plus en plus variés ?

Stéphanie Combes – *Division Méthodes appliquées de l'économétrie et de l'évaluation, Insee*

Résumés des interventions

Des Big Data pour la statistique publique ? Pourquoi et comment ?

Pauline Givord – *Division Méthodes appliquées de l'économétrie et de l'évaluation, Insee*

Le domaine des Big Data recouvre à la fois des sources de données originales ainsi que des méthodes nouvelles ou remises au goût du jour d'analyse de ces données (nowcasting, méthodes de classification, d'analyses de réseau, analyse textuelle et plus largement méthodes de machine learning). La statistique publique s'intéresse à ce domaine pour plusieurs raisons : compléter la production existante, raccourcir les délais de production, réduire les coûts de collecte ou encore assurer son positionnement parmi les pourvoyeurs d'information. Le projet « Données de caisse », porté par la DSDS et qui entrera bientôt en production, en constitue une illustration. De manière plus prospective, plusieurs projets expérimentaux sont menés depuis deux ans au sein du département des méthodes statistiques. Ces projets tentent d'évaluer le potentiel de données souvent évoquées comme intéressantes pour la statistique publique (données de téléphonie mobile, données issues de requêtes internet...), mais aussi de s'approprier les outils et techniques statistiques nécessaires pour exploiter ces données spécifiques. Cette présentation sera l'occasion de rappeler le contexte de ces expérimentations, d'en présenter les grandes lignes ainsi qu'un premier bilan.

Les outils du Big Data vu par un statisticien

Benjamin Sakarovitch – *Division Méthodes appliquées de l'économétrie et de l'évaluation, Insee*

L'exploitation de données massives peut rendre nécessaires de nouveaux supports informatiques, à la fois pour stocker et pour traiter des données de tailles gigantesques. Pour ce faire, des environnements logiciels adaptés ont été développés depuis une quinzaine d'années. Il existe aujourd'hui tout un panorama de solutions, la plus populaire étant Hadoop, qui fonctionne en parallélisant les tâches sur différents serveurs de stockage des données, à partir de la stratégie « Map-Reduce » qui sépare puis agrège les opérations. Plus récemment, l'arrivée du logiciel Spark a amélioré les performances des traitements sur de gros volumes. Le format des données massives est extrêmement varié, différentes technologies proposent donc différents stockage selon leur structure – ou leur absence de structure. Cela induit toute une série de langages pour les interroger. Leur évolution est constante et de nouvelles briques font régulièrement leur apparition. Ces solutions sont nouvelles, et encore peu utilisées à l'Insee.
Cette présentation n'a pas vocation a faire une revue exhaustive de ces différents outils, mais plutôt à présenter les premiers retours d'expérience, du point de vue d'un statisticien utilisateur, concernant les outils mobilisés dans le cadre des premières expérimentations menées.

Les méthodes de machine learning, un investissement utile pour la statistique publique

Stéphanie Combes – *Division Méthodes appliquées de l'économétrie et de l'évaluation, Insee*

Le « machine learning » n'est pas une discipline nouvelle, toutefois elle revient en avant avec les progrès technologiques permettant d'effectuer des calculs complexes en un temps raisonnable. Popularisée par les GAFA (Google, Apple, Facebook, Amazon) et souvent utilisée à des fins de prédiction (systèmes de recommandation par exemple), son champ d'application ne se limite cependant pas aux données massives. Les méthodes dites de machine learning se distinguent de la statistique et de l'économétrie traditionnelles par le peu d'hypothèses faites sur le processus générateur des données et par une spécification/calibration des modèles/algorithmes en grande partie automatisée. Cela peut s'avérer particulièrement pertinent lorsque l'on manque d'expertise sur les données manipulées (par exemple des données produites par un organisme privé telles que les Google Trends) ou lorsque le nombre de variables est trop important pour une sélection manuelle. La popularité croissante et l'accessibilité de ces méthodes, parfois fondées sur des approches statistiques anciennes telles que la régression logistique, ont attiré l'attention de certains instituts de statistique publique qui y voient des approches intéressantes dans des domaines variés : codification automatique, analyse conjoncturelle, identification d'anomalies et nettoyage de base de données, imputation, fusion de bases, extraction d'information dans les questions ouvertes d'enquête... Cet exposé décrira notamment comment ces techniques ont été mises en œuvre pour évaluer la pertinence de l'exploitation de données de téléphonie mobile agrégées pour une analyse territoriale menée à l'échelle nationale.

Comment exploiter des données aux formats de plus en plus variés ?

Stéphanie Combes – *Division Méthodes appliquées de l'économétrie et de l'évaluation, Insee*

La diversification des sources de données s'accompagne d'une diversification des formats : format texte ou données en réseau par exemple. Des méthodes spécifiques pour l'exploitation de ces données ont été développées depuis longtemps. Il s'agit, par exemple, d'utiliser la fréquence des termes dans des champs textuels pour identifier les thèmes abordés ou pour qualifier leur tonalité (positive, neutre ou négative), ou encore de caractériser les liens entre des unités pour les représenter sous la forme d'un réseau, au sein duquel on pourra détecter les nœuds influents par exemple. Si l'exploitation de ces données peut requérir un investissement conséquent, la généralisation de leur usage pour l'exploitation de nouvelles données, par exemple en provenance d'Internet, a permis la diffusion d'outils plus facilement accessibles que par le passé pour le non-spécialiste. Cette présentation fournira un premier aperçu pratique des principes de l'exploitation et de la modélisation de ces données atypiques, au travers d'exemples concrets.